- Motivation
- Past experience
- The LHConCRAY project
- System architecture and integration
- First preliminary performance data
- Outlook and Conclusions

ARC AS CRAY FRONT-END

SWISS EXPERIENCE

Gianfranco Sciacca, Sigve Haug

AEC - Laboratory for High Energy Physics, University of Bern, Switzerland

NorduGrid Conference 2016
Friday 2 June 2015, Pavol Jozef Safarik University, Kosice

MOTIVATION

The challenges of LHC computing for the next decade

- **The Worldwide LHC Computing Grid is made mainly of ah-hoc engineered computing sites**
- **The WLCG model doesn't scale for HL-LHC (beyond 2020)**
- Need (considerably) more computing for the same money

- Part of the solution is to consolidate LHC computing
  - Less but bigger sites world wide (operationally cheaper, better hw, etc)
  - General purpose HPC centres seem a good alternative to dedicated clusters
  - Lightweight operational approach can free up time for the crucial experiment support layer
  - The computing site does "computing", the matching to the experiment needs is for the experts
  - Cheaper than having experiment experts at all computing sites

- Some challenges arise
  - Processor architecture and/or OS might not always be suitable, complex software re-builds
  - Compliance with tight access rules
  - Application provisioning
  - Workload management
  - Data retrieval

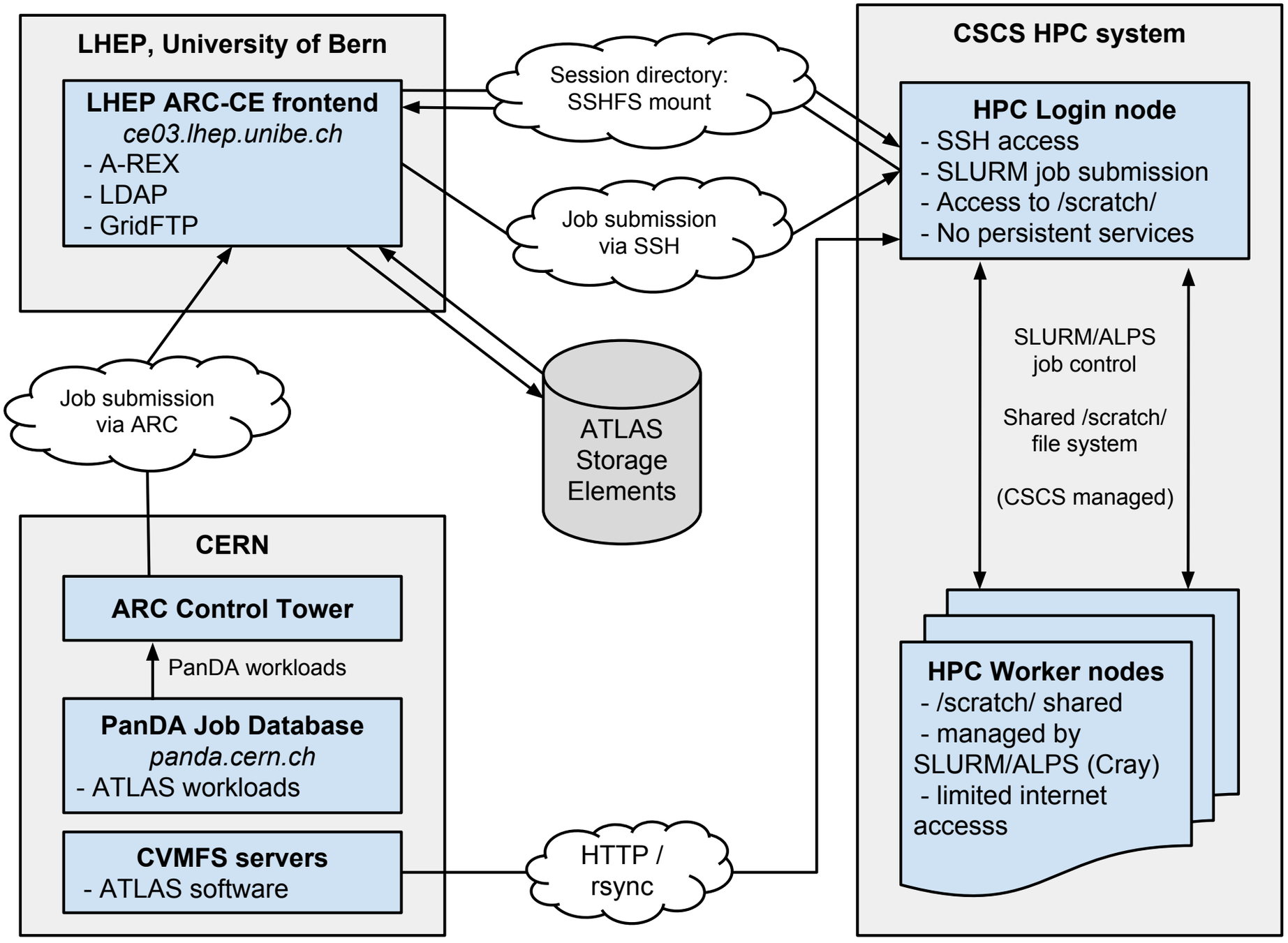
PAST EXPERIENCE

First approach to a Cray at CSCS, Lugano (2014/15)

System name	*Tödi*	*Piz Daint*	*Piz Dora*	*Monte Rosa*
Model	Cray XK7	Cray XC30	Cray XC40	Cray XE6
Description	Former CPU/GPU development and integration system.	Current flagship hybrid CPU/GPU system.	Flagship CPU-only system.	Former flagship CPU-only system.
Compute node configuration	• 16 core AMD Opteron CPU • 32 GB RAM • NVIDIA Tesla K20X GPU	• 8 core Intel Xeon CPU • 32 GB RAM • NVIDIA Tesla K20X GPU	• 2 x 12 core Intel Xeon CPUs • 64/128 GB RAM	• 2 x 16 core AMD Interlagos CPUs • 32 GB RAM
Number of compute nodes	272	5272	1256	1496
Total number of CPU cores	4352 + 272 GPUs	42176 + 5272 GPUs	30144	47872
Interconnect	Cray Gemini	Cray Aries	Cray Aries	Cray Gemini
Resource Manager / Scheduler	Cray SLURM / ALPS	Cray SLURM / ALPS	Cray SLURM / ALPS	Cray SLURM / ALPS

PAST EXPERIENCE

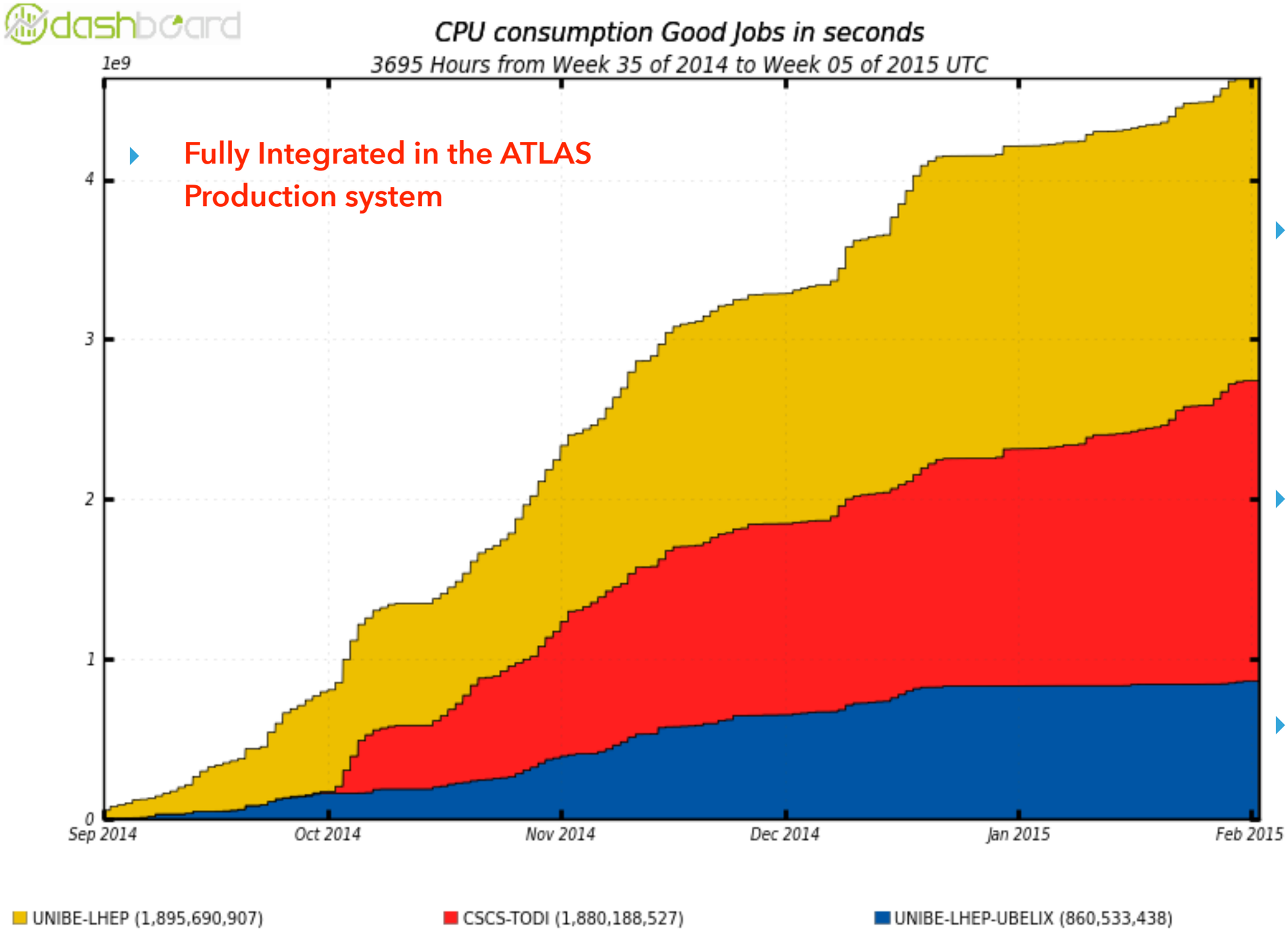
First approach to a Cray at CSCS, Lugano (2014/15)



- Master thesis work by Michael Hostettler, Universität Bern , 2015

PAST EXPERIENCE

First approach to a Cray at CSCS, Lugano (2014/15)



- Fully Integrated in the ATLAS Production system
- Tier-2 at Bern
- Cray
- Tier-3 at Bern

PAST EXPERIENCE

First approach to a Cray at CSCS, Lugano (2014/15): lessons learned

General

- **The ARC Compute Element is well suited to targeting remote compute systems in a non-intrusive way**
  - Abiding to the strict access policies typical of a HPC centre
  - Seamless integration of compute resources in the experiment workload management system
  - The data management system allows for transparent delivery and retrieval of data
  - Does not need to be physically located inside the computing centre
  - The modifications needed for this use case will be included in a future release of ARC

Performance

- **We were able to run un-modified binaries out of CVMFS on the Cray :-)**
- Pre-compiled gcc binaries out of CVMFS performed better than Cray re-compiled binaries (~30%)
- gcc + Cray recommended options brought only a marginal improvement (~5%) in processing time per event.
- From 10 to 100 compute nodes the job rate scaling is linear (as expected)
- Thread scaling is linear, with a slight offset due to the initialisation and finalisation steps
- Memory footprint for a ATLAS G4 job running on 16 cores is considerably lower that the 32GB available

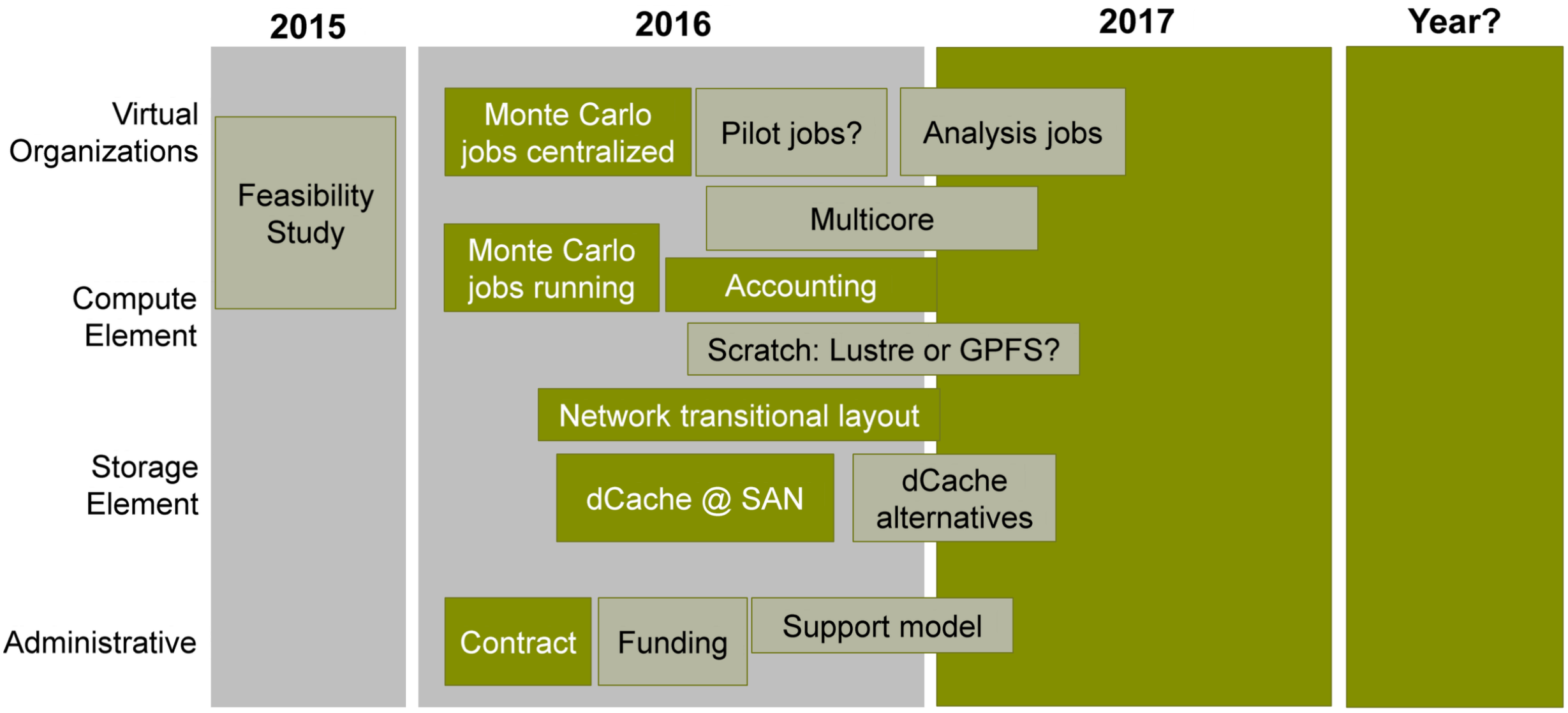
THE LHCONCRAY PROJECT: OVERVIEW

New approach to a Cray at CSCS, Lugano (2016)

- **Objective: be able to run all the Tier-2 Grid workloads on a share of general CSCS resources (big Cray systems, central storage, etc)**
- This includes:
  - All supported VOs: ATLAS, CMS, LHCb
  - All experiment workloads (pilot, production, analysis, monitoring, etc)
  - Fully integrated with the central factories (Crab, Panda, Dirac, etc)
  - Fully grid-aware (CE, SE, BDII, etc) and integrated into WLCG
  - "Green" for the VOs and the central EGI monitoring systems
- This poses many challenges
  - Difference in the OS and libraries
  - Supporting infrastructure such as CVMFS, scratch file system (10k files/job), network, etc.
  - Middleware readiness for ARC, Xrootd, GridFTP, SRM, Infosys, Accounting, etc.
  - VO readiness (they might need to adapt as well!) <= **VERY IMPORTANT**
  - Administrative, contractual and support changes
- **If the project is successful and shows financial advantage, the plan is to phase the current dedicated tier-2 systems out**

THE LHConCRAY PROJECT: OVERVIEW

New approach to a Cray at CSCS, Lugano (2016)



SYSTEM ARCHITCTURE AND INTEGRATION

Brisi: TDS for Piz Dora

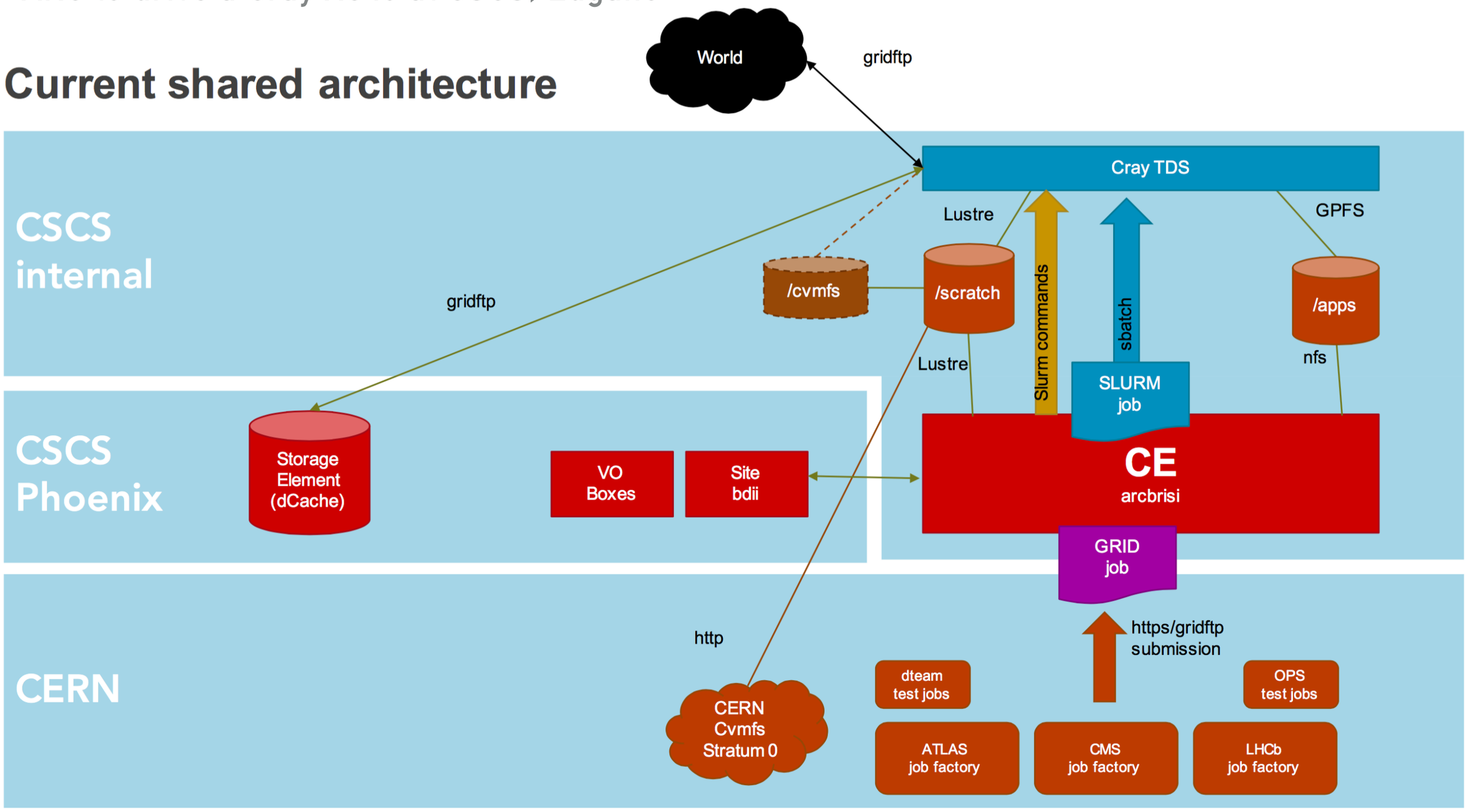
System name	*Tödi*	*Piz Daint*	*Piz Dora*	*Monte Rosa*
Model	Cray XK7	Cray XC30	Cray XC40	Cray XE6
Description	Former CPU/GPU development and integration system.	Current flagship hybrid CPU/GPU system.	Flagship CPU-only system.	Former flagship CPU-only system.
Compute node configuration	• 16 core AMD Opteron CPU • 32 GB RAM • NVIDIA Tesla K20X GPU	• 8 core Intel Xeon CPU • 32 GB RAM • NVIDIA Tesla K20X GPU	• 2 x 12 core Intel Xeon CPUs • 64/128 GB RAM	• 2 x 16 core AMD Interlagos CPUs • 32 GB RAM
Number of compute nodes	272	5272	1256	1496
Total number of CPU cores	4352 + 272 GPUs	42176 + 5272 GPUs	30144	47872
Interconnect	Cray Gemini	Cray Aries	Cray Aries	Cray Gemini
Resource Manager / Scheduler	Cray SLURM / ALPS	Cray SLURM / ALPS	Cray SLURM / ALPS	Cray SLURM / ALPS

- Native SLURM on Brisi

SYSTEM ARCHITCTURE AND INTEGRATION

ARC to drive a Cray XC40 at CSCS, Lugano

Current shared architecture



SYSTEM ARCHITCTURE AND INTEGRATION

LHConCRAY: Test integration phase ongoing with ARC CE

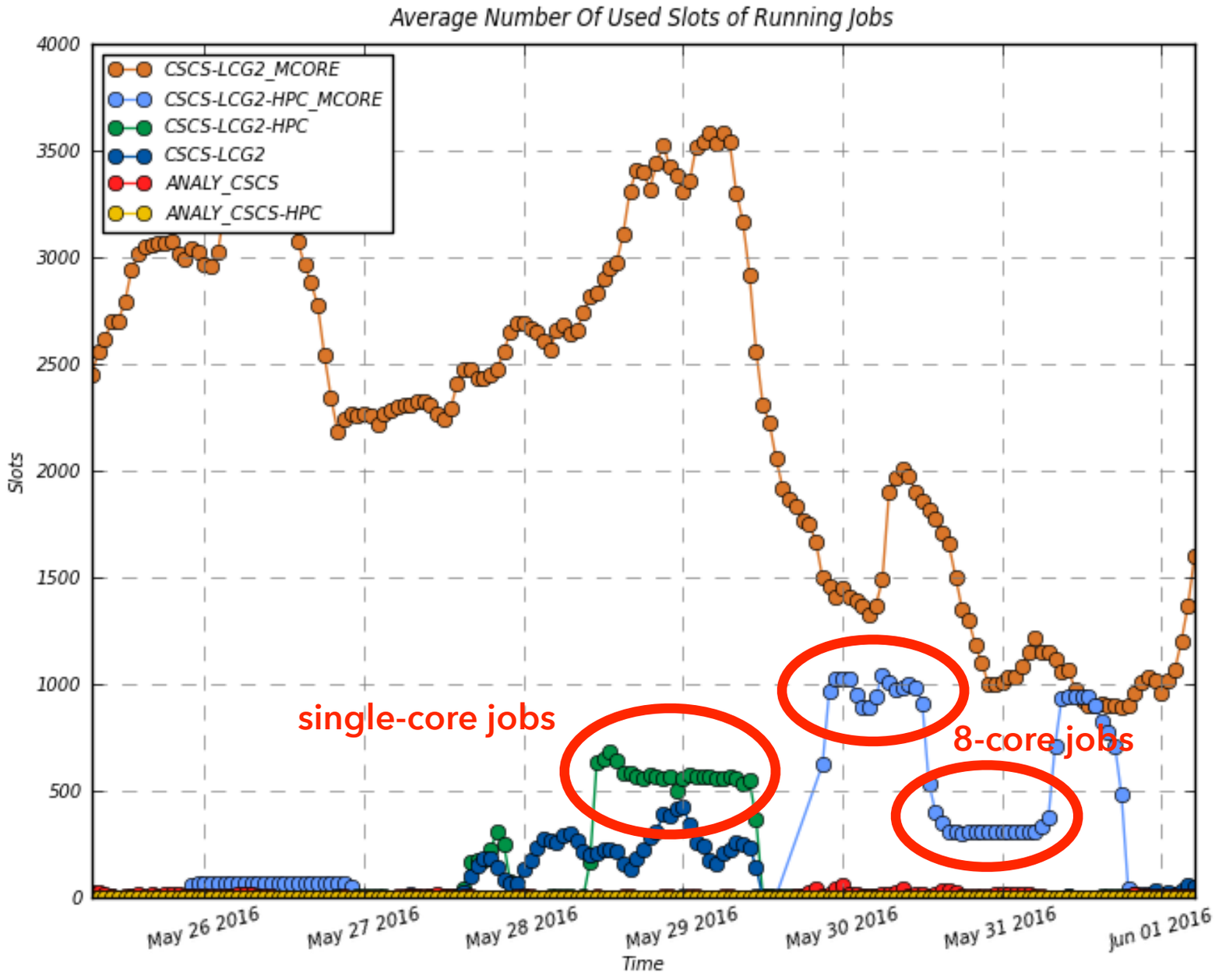
- Dedicated **ARC CE** in place (arcbrisi.cscs.ch): OPS, ATLAS, CMS and LHCb, submits directly to the LRMS.
- CVMFS served via NFS first. Bad performance, so switched to preloading the CernVM-FS Cache for all VOs.
- Compute: Cray TDS XC40 (Brisi)
  - 1 SLURM partition (wlcg) for all VOs.
  - **Broadwell** nodes with 64 HT-cores (Intel Xeon E5-2695 v4 @2.10GHz) and 128 RAM each.
  - Each node may be shared by concurrent jobs.
  - **HEPSPEC'06 rating** measured to be **13.39**/core. This is 18% higher when compared to the average rating of **11.46** for the current Tier-2 cluster (Phoenix)
  - Jobs run within Shifter containers [1]. The container itself is a CentOS 6.7 full image with mostly the same packages in Phoenix and configured accordingly.
- Integrated in the ATLAS PanDA workload management:
  - **CSCS-LCG2-HPC, CSCS-LCG2-HPC_MCORE, ANALY_CSCS-HPC**
- Also integrated in the CMS Crab factories and the LHCb Dirac factories

[1] http://www.nersc.gov/research-and-development/user-defined-images/

FIRST PRELIMINARY PERFORMANCE DATA

Performance with ATLAS HammerCloud stress tests

- Using the ATLAS HammerCloud testing framework to study performance for different workloads
- Workflow is designed for the test to be reproducible (same input data, 1k events and 8k events)

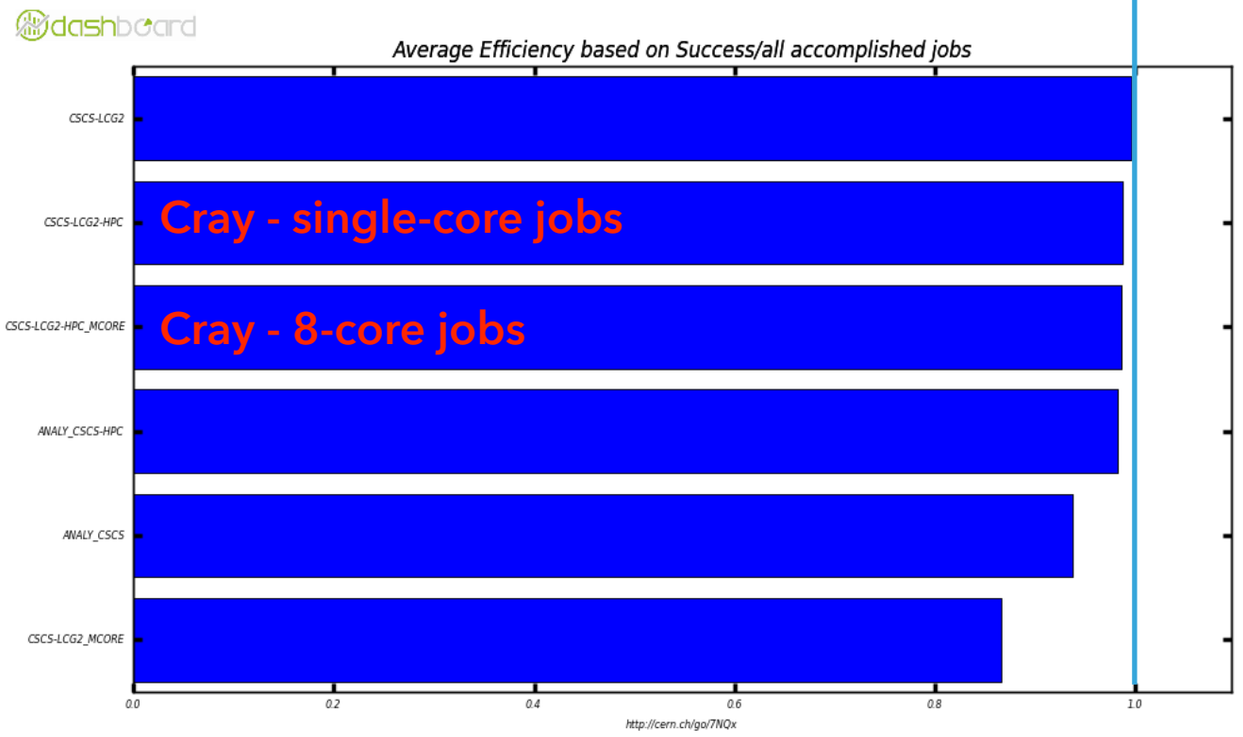


FIRST PRELIMINARY PERFORMANCE DATA

Performance with ATLAS HammerCloud stress tests: Job success rate

- CPU-bound workload: ATLAS detector simulation: **Job success rate**

Time ——————————>



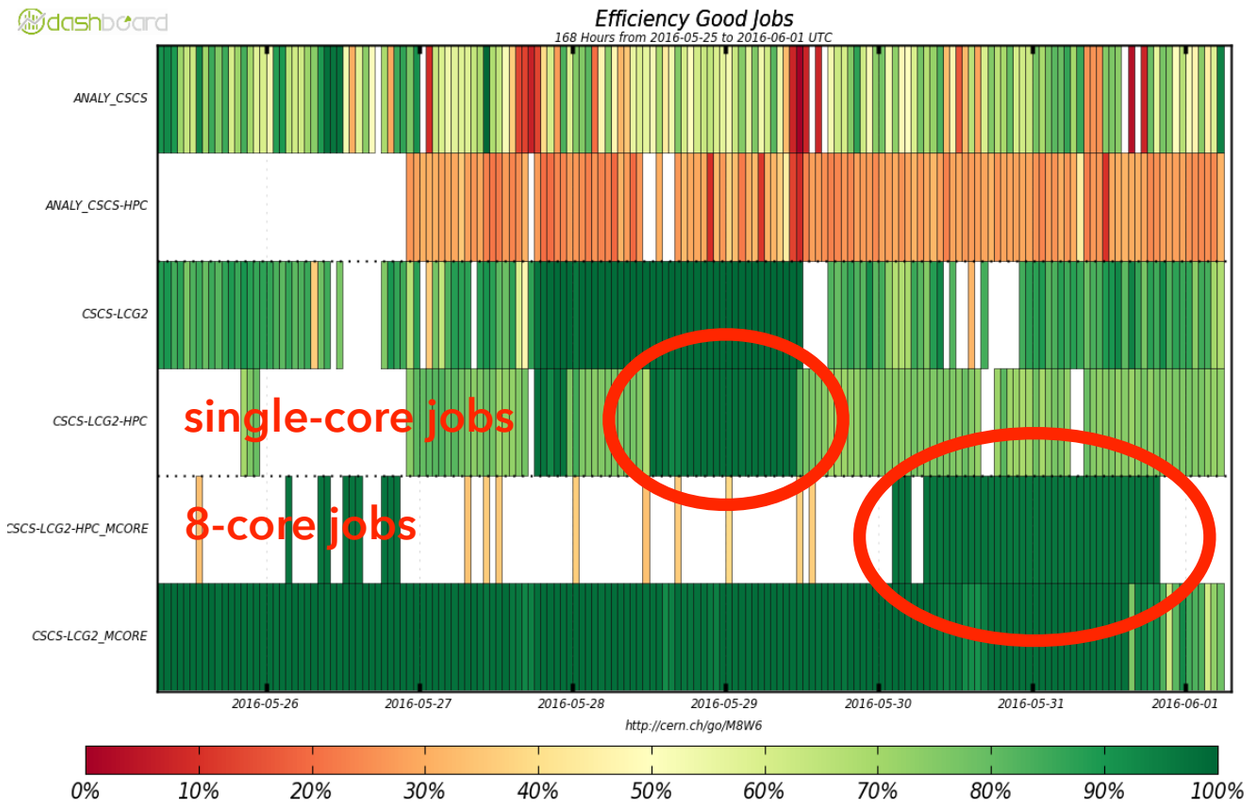
Job Efficiency ——————————>

1.0

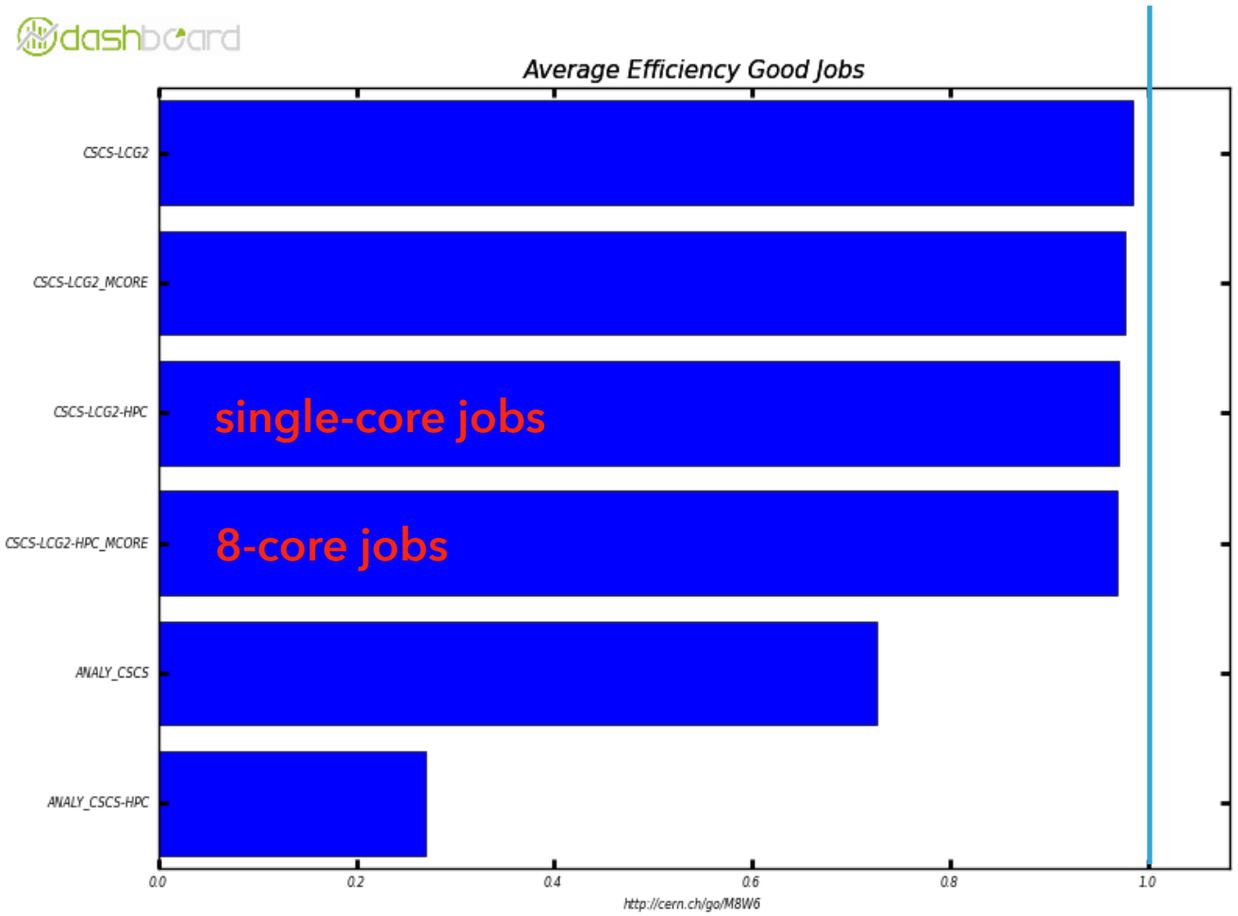
FIRST PRELIMINARY PERFORMANCE DATA

Performance with ATLAS HammerCloud stress tests: CPU efficiency

- CPU-bound workload: ATLAS detector simulation: **CPU efficiency**



Time ——————————>

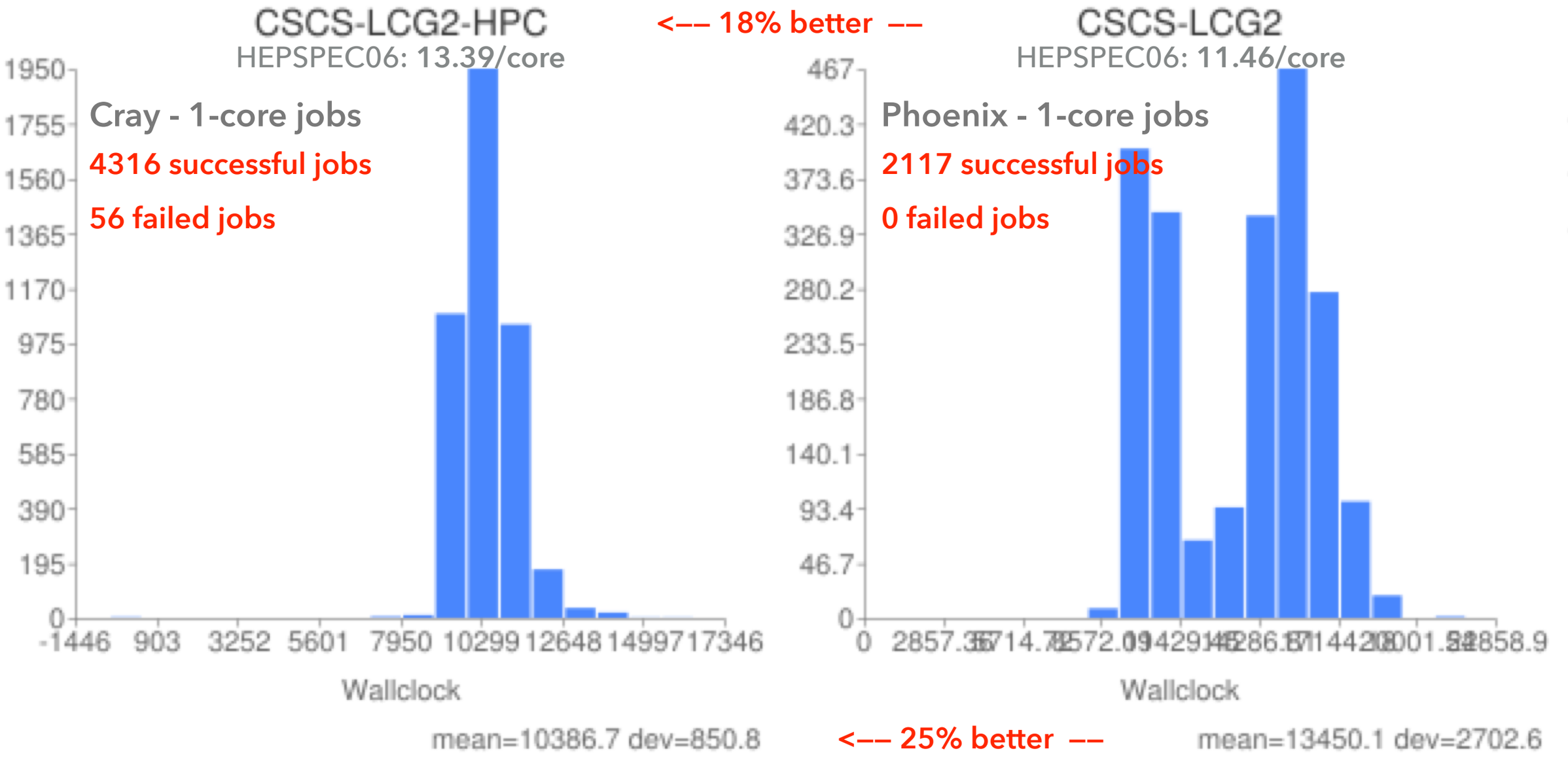


CPU/WT Efficiency ——————————> 1.0

FIRST PRELIMINARY PERFORMANCE DATA

Performance with ATLAS HammerCloud stress tests: WallClock

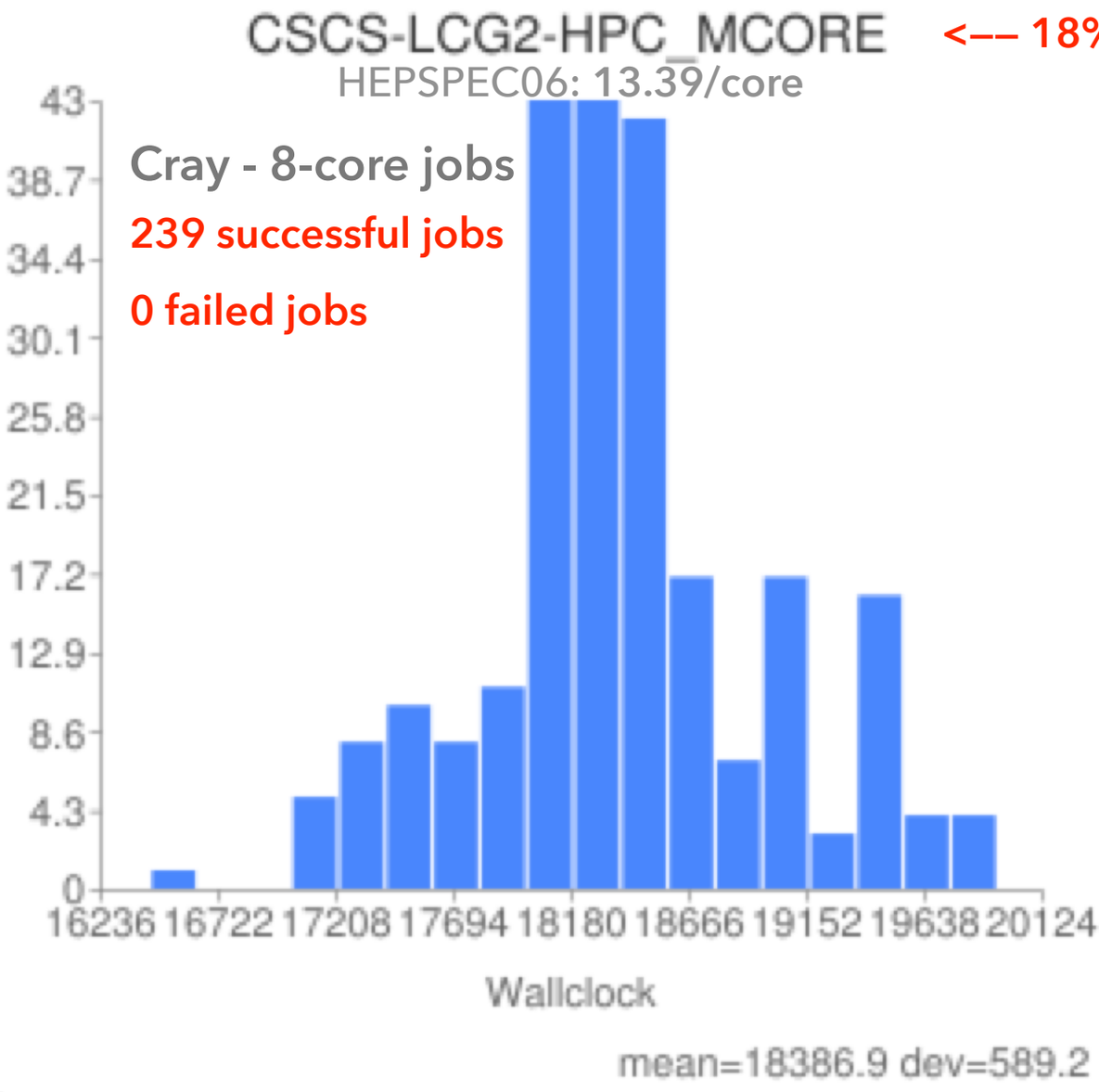
- CPU-bound workload: ATLAS detector simulation: **WallClock (single core)**



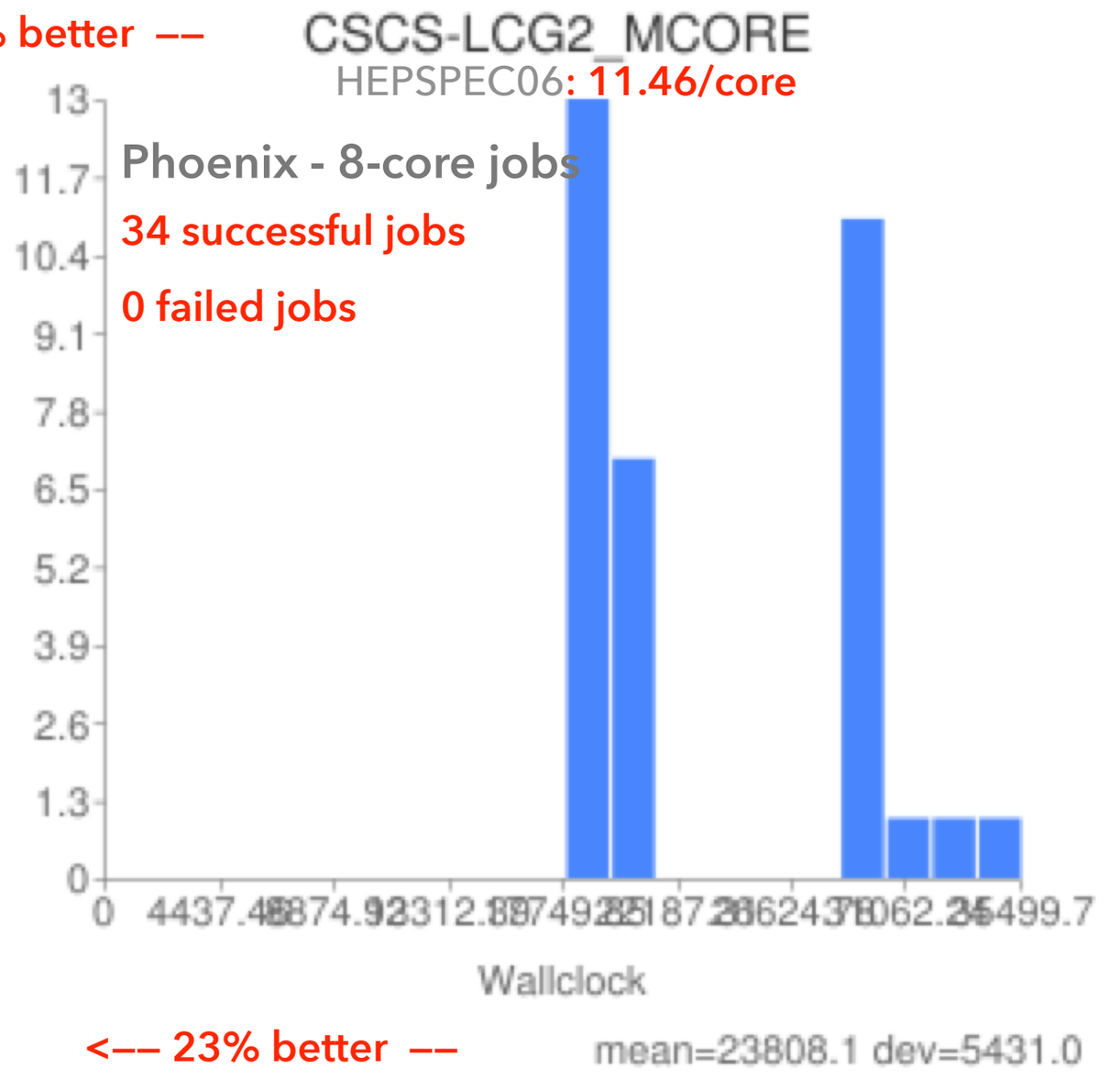
FIRST PRELIMINARY PERFORMANCE DATA

Performance with ATLAS HammerCloud stress tests: WallClock

- CPU-bound workload: ATLAS detector simulation: **WallClock (8 core)** (test 1/2)



<–– 18% better ––

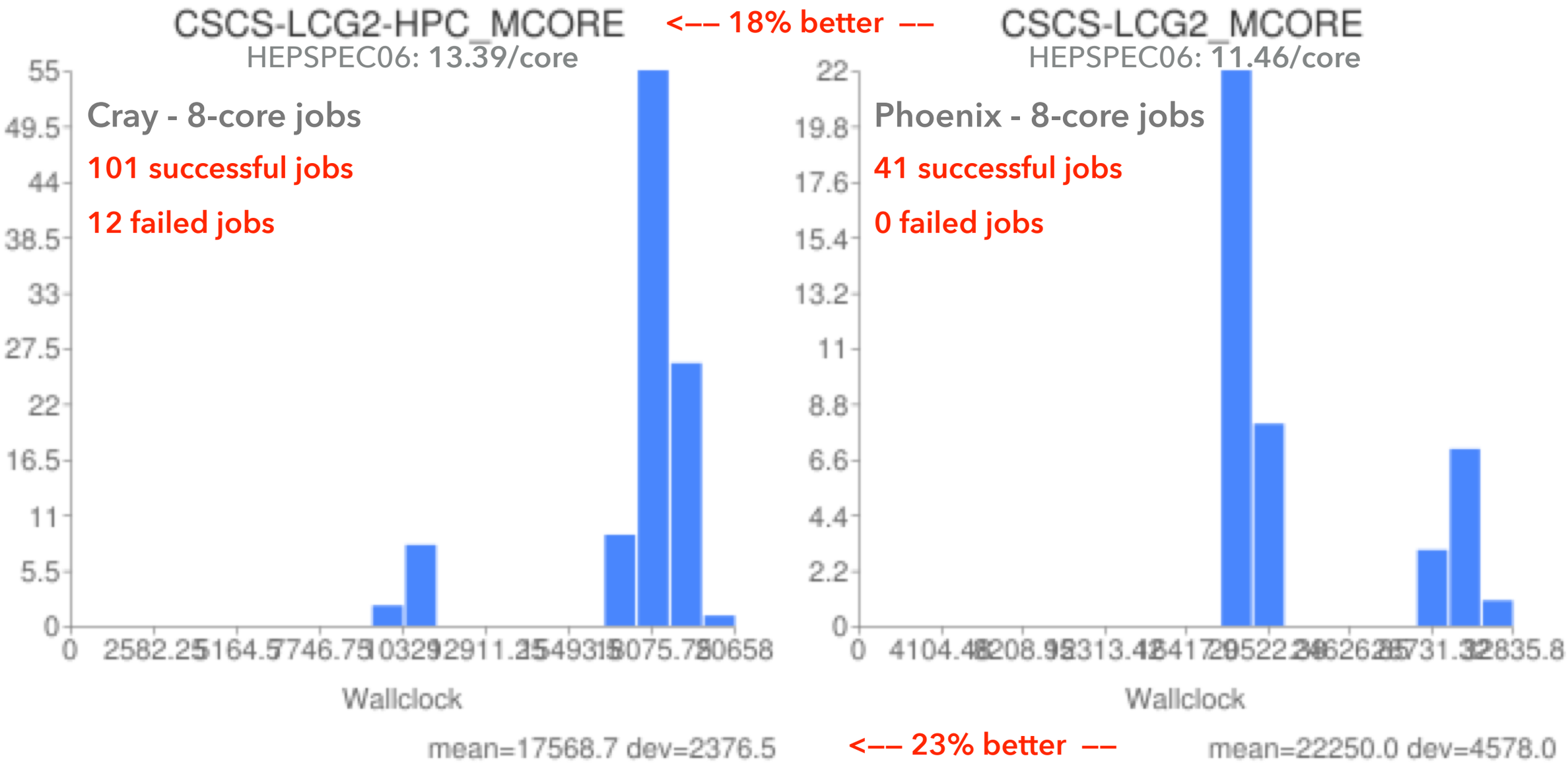


<–– 23% better ––

FIRST PRELIMINARY PERFORMANCE DATA

Performance with ATLAS HammerCloud stress tests: WallClock

- CPU-bound workload: ATLAS detector simulation: **WallClock (8 core)** (test 2/2)



OUTLOOK

Next steps

- Perform benchmarking with more workloads (ATLAS, I/O intensive), incl. user analysis jobs
- Ramp-up test: from zero to x*k cores, check ramp-up time (ARC+shared FS performance)
- Will likely need some tests beyond 1k-core scale
- Get real jobs to run : -)
- Slurm fair-share and accounting tuning (partially conflicting VO ecosystems)

- **Feed some of the information from above back to the cost study**

- Additional challenges arise from the different mode of operation/computing models of the 3 experiments: the ecosystem gets too complex, think hard here (or split? YES)

- Hard to judge the potential impact on/from other communities running on the machine

CONCLUSIONS

- We are trying to address some of the challenges of LHC computing for beyond the next 5 years
- The use of general purpose HPC systems is one of the possible ways forward

- **The ARC technology provides us with the right tool**
  - lightweight and non-invasive access to high-end computing resources
  - should contribute to bring down (considerably) operational costs

- Quite some experience gained doing real computations for ATLAS (with minimal support to it needed)

- We have setup the LHConCRAY project to explore the feasibility and financial advantages of this model
- Our provider doesn't like to de-couple pure computing infrastructure provisioning from support like we can do with ARC.

- **Right now the limitation is not the technology, rather politics**

- We aim at a decision after summer